Texto em revisão final; por favor, não citar sem autorização do autor
Parte do livro: COZMAN, Fabio, PLONSKI, Guilherme e NERI, Hugo (org.). *Inteligência Artificial e suas aplicações*. São Paulo: Instituto de Estudos Avançados da USP (em edição).

A tirania do acesso à informação.
Dominando a explosão digital de documentos textuais[1]

Alexandre Moreli
Instituto de Relações Internacionais/USP
Faculty Researcher – History Lab (www.history-lab.org)

Introdução

Em um tempo em que mais da metade da população mundial acessa a internet deixando registro de suas comunicações em trocas de e-mails, de seus pensamentos ou crenças em *tweets* ou de sua vida familiar compartilhando fotos online, dentre inúmeras angústias, surge uma menos evidente hoje: como pesquisar essas diferentes experiências no futuro? Como dominar a complexidade da construção de novas narrativas sobre o passado diante de tantos traços da experiência humana agora preservados?

Mais do que nunca, uma reflexão a respeito da passagem humana pelo tempo e pelo espaço tem o potencial de revelar contextos labirínticos sobre transformações e

[1] Texto base da apresentação "Arquivos e narrativas sem fronteiras" no evento *Inteligência Artificial e suas Aplicações: Avanços e Tendências*, de 25 de junho de 2019, da série *Strategic Workshops* do Instituto de Estudos Avançados da USP (Acesso online http://www.iea.usp.br/midiateca/video/videos-2019/inteligencia-artificial-e-suas-aplicacoes-avancos-e-tendencias-parte-1-de-2). Agradeço ao apoio dos bolsistas de Iniciação Científica da Universidade de São Paulo Maria Victoria Villela, Thales Rodriguez e Lucca Rocha na construção deste texto. Agradeço, também, os comentários de Marcos Lopes, Adriana Schor, Nelly De Freitas e Luciana Heymann e a colaboração dos colegas do *History Lab*, do Laboratório de História Global e das Relações Internacionais do Instituto de Relações Internacionais da USP (Labmundi-IRI) e do Centro de Inteligência Artificial da USP.

permanências. Será possível, entretanto, dominar complexidades reveladas de uma forma sem precedentes?

Até agora, alcançar e entender tal enredamento era, na verdade, vencer a erosão sobre os vestígios do tempo pregresso. Durante décadas, desde quando as ciências históricas passaram a tornar o trato das fontes como intrínseco à pesquisa, chegou-se até mesmo a considerar o passado como um tirano do historiador. Esse último não seria um ser livre, dizia Marc Bloch, por ser impedido de conhecer qualquer coisa a não ser o que o próprio passado lhe revelasse[2]. Com o advento da computação e do universo digital, ou seja, de uma nova e muito maior capacidade de armazenamento, processamento e preservação dos mesmos vestígios, imaginava-se que tal tirania iria desaparecer. Com a explosão digital de documentos textuais, entretanto, ela parece ter apenas se transformado.

Todo aquele interessado em entender a sociedade tem enfrentado um aumento sem precedentes no número de registros da atividade humana, mas também dificuldades em entender as novas tipologias desses registros e suas formas de preservação. Para além das experiências individuais e privadas digitais, governos e instituições também parecem agora existir sobretudo virtualmente, abandonando os registros físicos de suas atividades e alterando o que permanecerá para a posteridade. O Arquivo Nacional dos Estados Unidos, por exemplo, já anunciou um plano estratégico em que prevê não mais receber arquivos em papel a partir de 2023[3].

O historiador sempre teve que se preocupar em dominar a densidade do passado que, longe de deixar traços lineares, é formado por uma sedimentação de elementos muitas vezes contraditórios. Diante desse dever de reflexão, a outrora tirania da escassez manifesta-se, hoje, na explosão do número de vestígios, nas consequentes dificuldades em organizá-los, na impossibilidade de examiná-los em conjunto manualmente dadas as novas dimensões de escala e também na contínua diversificação de perspectivas sobre o olhar para o passado. Essa combinação levanta questões práticas e metodológicas que interessam todos, para muito além do oficio do historiador.

---

[2] BLOCH, Marc. **Apologia da História Ou o ofício de historiador**. Rio de Janeiro: Zahar, 2002 (1949), p. 75.

[3] LAWRENCE, Kerri. *National Archives New*, 23/08/18. Disponível em <https://www.archives.gov/news/articles/leaders-share-national-archives-vision-for-a-digital-future>. Acesso em: 03 Set. 2019.

Este capítulo busca fazer uma avaliação do que William McAllister chamou de "Big Bang documental"[4], sobretudo através do reconhecimento de como registros textuais criados, armazenados e acessados em formatos digitais desafiam pesquisadores das Ciências Humanas, em geral, e historiadores, em particular, em seus ofícios. Ao mesmo tempo, preocupa-se em identificar quais métodos (como Aprendizado de Máquina e Processamento de Linguagem Natural) e parâmetros podem ser aplicados ou desenvolvidos para um tempo em que suportes como *Facebook* e *Twitter* tornam-se tanto registros da experiência humana como arquivos históricos[5]. Finalmente, com uma preocupação em transpor esse conhecimento para outras áreas de aplicação, como jornalismo, análise de redes sociais ou de mercado consumidor, procura refletir sobre como contornar os desafios tecnológicos para tornar o trabalho com documentos textuais em larga escala e não estruturados acessível a pesquisadores, formuladores de políticas públicas ou quaisquer interessados de uma forma abrangente e útil.

O horizonte do arquivo infinito e suas implicações

Para alguém trabalhando com gestão da informação pública, ou ainda formado ou realizando pesquisas nas Ciências Humanas até o final do século XX, o impacto da digitalização dos processos sociais das últimas duas décadas certamente representa uma notória transformação. Esse fenômeno acarreta, simultaneamente, a digitalização de fontes de pesquisa e provoca um impacto sem precedentes na massa documental acumulada. Não se faz inédito hoje, entretanto, o recurso a sistemas automáticos e inteligentes de análise para enfrentar esse tipo de desafio. O potencial da Estatística e da Computação, por exemplo, já é conhecido de longa data.

Historiadores econômicos, nesse sentido, há muito, demonstram um interesse por métodos quantitativos e estatísticos, nomeando fontes como "dados" e as analisando com

[4] MCALLISTER, William, The Documentary Big Bang, the Digital Records Revolution, and the Future of the Historical Profession, **Passport**, n. 41, vol. 2, 2010, p. 12.

[5] Enquanto para o *Facebook* ainda não existe acesso franco ao seu histórico de conteúdo, lembrando que seus usuários trocam diariamente mais de 300 milhões de mensagens, absolutamente todos os *twiters* enviados entre 2006 e 2017 encontram-se arquivados no *Twitter Archive* da Biblioteca do Congresso dos Estados Unidos, constituindo fontes históricas para os pesquisadores (Library of Congress. *Update on the Twitter Archive at the Library of Congress*. Dezembro de 2017. Disponível em <https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf>. Acesso em: 22 Ago. 2019).

técnicas inovadoras de modelagem. No início do trabalho com tal metodologia, a euforia fora tanta que proliferaram discursos pregando uma renovação completa na capacitação dos estudantes, defendendo que as formações deveriam, na verdade, concentrar-se em Estatística e Programação. Tudo isso, já na década de 1950. Como lembram David Allen e Matthew Connelly, membros desse movimento chegavam a afirmar que "os métodos quantitativos iriam dominar a História, transformando-a de arte em ciência e livrando a profissão da dissimulação ideológica"[6]. Logo percebeu-se, entretanto, que nem todas as perguntas poderiam ser respondidas quantitativamente, ainda que tradições como da "Cliometria" ou da "História Serial" (experiências de pesquisas preferindo análises históricas quantitativas) tenham consolidado sua legitimidade[7].

No século XXI, tanto quanto o desenvolvimento de novas técnicas de pesquisa e a evolução das metodologias, a explosão de dados seriais e de documentos textuais não estruturados está levando imperiosamente as Ciências Humanas a um semelhante novo momento de reflexão sobre os clássicos paradigmas da ciência. O fato de esse momento poder ser definido como o das "Humanidades Digitais" ainda permanece como um debate em aberto, sobretudo quanto aos seus contornos e conteúdo[8]. Certo é o impacto de uma nova escala de produção e preservação de informação experimentada neste início de século.

Particularmente quanto à informação produzida por autoridades públicas, reconhece-se duas importantes consequências: primeiro, as dificuldades de sua gestão e

[6] ALLEN, David e CONNELLY, Matthew. Diplomatic history after the big bang: using computational methods to explore the infinite archive, COSTIGLIOLA, Frank e HOGAN, Michael J. **Explaining the History of American Foreign Relations**. Nova York: Cambridge University Press, 2016, p. 76.

[7] Entre a *Cliometria* e a *História Serial*, a disciplina renovou-se em meados do século XX. Para um histórico, ver: NORTH, Douglass C. Cliometrics – 40 Years Later, **The American Economic Review**, vol.87, n.2, 1997, p.412- 414 e FLORENTINO, Manolo e FRAGOSO, João, A História Econômica: Balanço e Perspectivas Recentes, CARDOSO, Ciro Flamarion (org.). **Domínios da História: ensaios de teoria e metodologia**. Rio de Janeiro: Editora Campus, 1997, p. 27-43. Interessante notar que Ciro Flamarion Cardoso, já em meados dos anos 1970 no Brasil, refletia sobre um uso mais ambicioso do computador no ofício do historiador (CARDOSO, Ciro Flamarion, O uso da computação em História, CARDOSO, Ciro Flamarion. **Os Métodos da história**. Rio de Janeiro: Edições Graal, 1979 [1976], p. 503-510).

[8] Para algumas sínteses sobre o debate no mundo lusófono, ver: ALVES, Daniel, As Humanidades Digitais como uma comunidade de práticas dentro do formalismo académico: dos exemplos internacionais ao caso português, **Ler História**, n. 69, 2016, p. 91-103 e PIMENTA, Ricardo M., Das iniciativas em Humanidades Digitais e suas materialidades: relato de um laboratório em construção contínua, **Memória e Informação**, v. 3, n. 1, 2019, p. 1-14.

de sua preservação pelo próprio Estado e, segundo, as dificuldades de análise por pesquisadores.

Em um caso emblemático, o *Information Security Oversight Office*, autoridade federal estadunidense que vela pela proteção e pelo acesso a informações produzidas pelo Estado, tem frequentemente emitido alertas quanto à gestão de documentos secretos e ultrassecretos. Para além de levantar números tão impressionantes como o de 50 milhões de documentos sigilosos sendo produzidos anualmente pelo governo dos Estados Unidos[9], aponta para a dificuldade de setores da administração americana e do próprio Arquivo Nacional do país (o *National Archives and Records Administration*-NARA) em rever as classificações e liberar acesso a toda a documentação já produzida.

Ainda que mais de 1 bilhão de documentos tenham sido desclassificados nas últimas três décadas nos Estados Unidos, a falta de meios materiais e restrições orçamentárias tem atrasado as revisões manuais dos documentos com classificação mais sensível ou os pedidos de liberação feitos diretamente por cidadãos através da Lei de Acesso à Informação do país (*Freedom of Information Act*-FOIA)[10]. Há hoje pedidos feitos através do FOIA junto ao NARA com mais de 25 anos[11]!

O *Public Interest Declassification Board*, comitê criado pelo Congresso dos Estados Unidos para promover o maior acesso possível à documentação, já alertou a administração americana de que "é preciso haver uma conscientização e um acordo de que a prática atual de ter uma, duas ou mais pessoas realizando uma laboriosa avaliação de desclassificação, página por página, para cada registro em análise é uma prática insustentável"[12]. Dentre as seis recomendações mais urgentes feitas pelo órgão, pode-se ler a de que "o governo deve exigir que as agências desenvolvam e usem novas

[9] ISOO, *2017 Report to the President*, p. 44. Disponível em <https://www.archives.gov/files/isoo/reports/2017-annual-report.pdf>. Acesso em: 23 Ago. 2019.

[10] Public Interest Declassification Board. **Setting priorities: an essential step in transforming declassification**. Dezembro, 2014, p. 15. Para a legislação sobre desclassificação automática nos Estados Unidos, considerar a *Executive Order* 12958 (*Classified National Security Information*) de 1995 e a *Executive Order* 13526 de 2009.

[11] HARPER, Lauren, JONES, Nate, BLANTON, Tom e REID, Tena-lesly. 25-Year-Old FOIA Request Confirms FOIA Delays Continue Unabated, **National Security Archive**, 08/03/19. Disponível em [https://nsarchive.gwu.edu/foia-audit/foia/2019-03-08/25-year-old-foia-request-confirms-foia-delays-continue-unabated]. Acesso em: 23 Ago. 2019.

[12] Public Interest Declassification Board. Op. cit., p. 15.

tecnologias para auxiliar e melhorar a revisão de desclassificação" [13]. Essa recomendação se torna ainda mais válida quando se sabe que, hoje, o Departamento de Estado dos EUA, por exemplo, produz 2 bilhões de e-mails por ano, ou que uma única agência de segurança nos Estados Unidos produz, a cada 18 meses, cerca de 1 *petabyte* de informação classificada, material suficiente para preencher 20 milhões de gavetas caso impresso.[14]. O NARA estima que, sem novas tecnologias para acelerar o processo, que se trata essencialmente de leitura, análise, interpretação e tomada de decisão sobre se e quando textos secretos devem ser liberados para acesso público, somente esse último montante mencionado necessitaria de dois milhões de funcionários por ano para passar pelo processo de desclassificação enquanto, na realidade, há apenas 41 arquivistas trabalhando para revisar registros de todo o governo federal... uma página por vez, manualmente[15]! Ademais de sua gravidade, trata-se de apenas um entre vários desafios à gestão da informação que órgãos responsáveis pela transparência pública começam a enfrentar.

Ainda que, no Brasil, não existam os mesmos recursos ou a mesma estrutura independente de acompanhamento do impacto da explosão digital de documentos textuais sobre sua preservação, diversas instituições mantenedoras de arquivos, inclusive o Arquivo Nacional, têm também sentido suas consequências.

De acordo com seu Relatório de Atividades de 2018, o Arquivo Nacional conserva hoje mais de 60 quilômetros de documentos textuais (lembrando que a mensuração é feita considerando cada folha de papel como enfileirada em posição vertical face a face). O acervo digital total, entretanto, ocupa apenas 494 terabytes[16], menos da metade do que um órgão de segurança do governo americano produz em pouco mais de um ano. Desse material, apenas uma pequena parcela está disponível online através do Sistema de

[13] Ibidem, p. 15.

[14] Public Interest Declassification Board, Using Technology to Improve Classification and Declassification, **The Blog of the Public Interest Declassification Board hosted by the National Archives**, 14/11/11. Disponível em [https://transforming-classification.blogs.archives.gov/2011/03/14/using-technology-to-improve-classification-and-declassification]. Acesso em: 23 Ago. 2019.

[15] CONNELLY, Matthew e IMMERMAN, Richard H. What Hillary Clinton's Emails Really Reveal, **New York Times**, 04/03/15. Disponível em [https://www.nytimes.com/2015/03/04/opinion/what-hillary-clintons-emails-really-reveal.html]. Acesso em: 23 Ago. 2019.

[16] Arquivo Nacional. Relatório de Atividades, 2018, p. 22. Disponível em <http://arquivonacional.gov.br/images/ASCOM/Relatorio_atividades_AN_2018a.pdf>. Acesso em 29 Out. 2019.

Informações do Arquivo Nacional (SIAN). Com um ritmo lento, em 2017 (último dado disponível) houve a digitalização de apenas 120.154 documentos, que se somaram aos quase 3 milhões de itens digitais existentes. Desses, entretanto, apenas 178.168 estavam disponíveis para acesso no SIAN[17]. Um número maior de documentos já digitalizados não pode ainda ser disponibilizado, pois necessita de tratamento por parte das equipes do Arquivo Nacional, como para a criação de descritores, que ainda é feita manualmente.

A evolução da procura por documentos nesse suporte, entretanto, mostra uma fortíssima demanda do público. Os acessos online passaram de 232.223 em 2016 para 1.157.209 em 2017, chegando a 2.093.354 em 2018[18]. Pesquisas mensais sobre a satisfação dos usuários realizadas pelo próprio Arquivo Nacional confirmam o interesse, mas também revelam que, dentre as críticas aos serviços prestados, constam a dificuldade de consulta ao SIAN e o baixo índice de digitalização do acervo[19].

Apesar da publicação do Decreto nº 8.539, em 2015, que estabelece o Processo Eletrônico Nacional no âmbito da Administração Pública no país, determinando a adoção de ações que garantam o acesso, o uso contínuo e a preservação a longo prazo dos documentos digitais, parece ser dramático o caso brasileiro. De todas as coleções preservadas em formato digital no Arquivo Nacional, de fato, apenas cerca de 1% nasceu digitalmente (as referentes às já extintas Comissão Nacional da Verdade e Autoridade Olímpica). Esse número indica que os documentos nascidos digitalmente não estão chegando ao órgão responsável pela guarda permanente, revelando uma fatalidade que ainda não mostrou sua extensão para todos os interessados em ter acesso à informação pública registrada desde a virada do século XXI[20].

Nesse mesmo sentido, conforme a sociedade se volta para as mídias sociais como método substancial de comunicação e expressão criativa, nota-se uma justaposição do

[17] Arquivo Nacional. Relatório Síntese do Exercício 2017, 2017, p. 3. Disponível em <http://arquivonacional.gov.br/images/Relatorio_de_gestao/Relatorio_Sintese_AN_2017_final.pdf>. Acesso em: 29 Out. 2019.

[18] Arquivo Nacional. Relatório de Atividades, 2018, p. 28. Disponível em <http://arquivonacional.gov.br/images/ASCOM/Relatorio_atividades_AN_2018a.pdf>. Acesso em: 29 Out. 2019.

[19] Arquivo Nacional. Relatório de Atividades, 2018, p. 89. Disponível em <http://arquivonacional.gov.br/images/ASCOM/Relatorio_atividades_AN_2018a.pdf>. Acesso em: 29 Out. 2019.

[20] Informações colhidas junto a funcionários do Arquivo Nacional pelo autor em maio de 2019.

digital (muitas vezes superação) quanto a manifestações antes registradas em cartas, periódicos ou outros suportes físicos. O neurocientista e filósofo Georges Chapouthier e o engenheiro da computação Frédéric Kaplan chegam mesmo a comparar essas novas memórias computacionais ao surgimento de técnicas explicitamente destinadas a conservar e a transmitir informações que impactaram profundamente a humanidade no passado, "como a linguagem oral, as pinturas rupestres, as escritas cuneiformes, os alfabetos e a impressão"[21]. Essas novas técnicas de arquivamento e preservação de conteúdo das plataformas de mídias sociais, ao mesmo tempo em que permitem que pesquisadores, no futuro, tenham acesso a uma visão mais completa das normas, diálogos, tendências e eventos culturais contemporâneos, reforçam a angústia tirânica da superabundância que começa a abalar os ofícios de pesquisa[22].

Repensar os arquivos e ler à distância

Difícil imaginar, entretanto, que existam soluções milagrosas, imediatas e triviais a serem rapidamente importadas da Ciência da Computação para as Humanidades ou que todos os interessados em analisar tais registros textuais terão que se transformar em programadores, ainda que se mostre importante ter noções sobre como códigos funcionam, sobre como arquivos digitais são armazenados e sobre a capacidade e as limitações de intervenção humana em cada um desse processos. As já mencionadas novas dimensões de escalas de textos disponibilizados alimentam de uma forma diversa a preocupação sobre como lidar com conjuntos de documentos textuais desorganizados e de diferentes tipologias, sem mencionar a necessidade de se medir as razões das ausências de documentos, dos descartes ou de trechos corrompidos em registros digitais para dar mais sentido ao todo.

Tais inquietações provocam um exercício de reflexão em duas frentes. A primeira, na Arquivologia, quanto à concepção de arquivo e, a segunda, na História, quanto a metodologias e ferramentas de pesquisa. Para a primeira, interessante o argumento

[21] CHAPOUTHIER, Georges e KAPLAN, Frédérik. **L'homme, l'animal et la machine**. Paris, CNRS Éditions, 2011, p. 29.

[22] CONNELLY, Matthew. The Next Thirty Years of International Relations Research. New Topics, New Methods, and the Challenge of Big Data, **Les cahiers Irice**, vol. 2, n. 14, 2015, p. 85-97.

adiantado por Michael Moss, David Thomas e Tim Gollins, de que arquivistas devem agora mudar suas perspectivas passando a considerar arquivos (que contarão com parcelas equilibradas em número de documentos textuais, de registros sonoros e de imagens no futuro) como "coleções de dados a serem minerados e não de textos a serem lidos"[23]. Para esses três especialistas da área, os arquivistas devem imperativamente observar que os historiadores estão reconsiderando seus métodos de pesquisa devido à explosão no número de registros administrativos digitais criados por órgãos públicos e de sua conjugação com veículos de comunicação online e com as mídias sociais.

O desafio seria entender as razões pelas quais recursos como ordem original ou hierarquia de funções no momento da produção tornaram-se limitados ou nulos para serem utilizados na compreensão de um conjunto documental. Não se tratam, portanto, apenas de questões ligadas à gestão de uma nova escala de registros e das consequentes dificuldades em tomar decisões sobre o que selecionar para preservação. Se uma determinada instituição arquivística pública costumava receber documentos de órgãos oficiais, cabendo ao pesquisador procurar outras instituições mantenedoras de arquivo quando desejasse cruzar fontes (como hemerotecas, arquivos de empresas ou de organizações não-governamentais), a forma como hoje as plataformas digitais produzem ou captam informações e dados estimula fortemente uma reflexão sobre preservação e acesso de uma forma mais ampla.

Tomemos o exemplo dos atentados terroristas ocorridos em Paris em 2015. Certamente, o Arquivo Diplomático e o Arquivo Nacional franceses receberão imensas quantidades de registros digitais quando chegar o momento da preservação permanente dos documentos públicos criados no momento e em razão dos atentados, o que, por si só, constituirá um desafio arquivístico por todo o exposto até agora. Entretanto, soma-se a essa questão os demais registros sincrônicos ao evento que já foram preservados por outras instituições, inclusive não governamentais como o *Internet Archive*[24]. Para além da excepcionalidade em razão da natureza impiedosa dos fatos, os atentados, mas também as reações a eles em termos de preservação, podem ser tomados como uma experiência para o que Moss, Thomas e Gollins propõem como discussão. De fato, notou-se uma

[23] MOSS, Michael, THOMAS, David e GOLLINS, Tim. The Reconfiguration of the Archive as Data to Be Mined, **Archivaria – The Journal of the Association of Canadian Archivists**, n. 86, 2018, p. 118.

[24] The Internet Archive. Disponível em <https://archive.org/ >. Acessado em: 12 Nov. 2019.

preocupação de arquivistas em como preservar as reações ao atentado no mundo real e no virtual.

No mundo real, diante de quase 8000 desenhos de criança, cartas, poemas, origamis e outros tributos espontaneamente depositados em frente ao local dos ataques de 13 de novembro de 2015, o Arquivo da Cidade de Paris, com o apoio dos serviços públicos de limpeza e administração, decidiu coletar, armazenar e digitalizar o material, que já se encontra disponível para consulta online[25].

Imagem 1: Documento identificado como "Desenho, Texto, Dobradura, Colagem e Objeto em Papel", coletado na praça Bataclan, em Paris (um dos locais do ataque terrorista) em 12 de janeiro de 2016, digitalizado e disponível no site do Arquivo Municipal de Paris[26]

[25] Archives. Le site des archives de Paris. Hommages aux victimes des attentats de 2015. Disponível em <http://archives.paris.fr/r/137/hommages-aux-victimes-des-attentats-de-2015/ >. Acessado em 12 Nov. 2019.

[26] Ibidem.

Para além dessas coleções, a Biblioteca Nacional e o Instituto do Audiovisual, também da França, lançaram em urgência iniciativas de captação e arquivamento em tempo real de reações aos atentados no *Twitter* e em outras plataformas da internet. Particularmente nesse caso do mundo virtual, surgem questões sobre como imaginar processos de captação e arquivamento em tempo real, sobre como transformar esse tipo de arquivo em corpus para a pesquisa[27], sobre o que se pode extrair dessas fontes nascidas digitalmente para analisar os acontecimentos no plano local, nacional e internacional, mas também para entender a participação dos internautas, as formas de expressão online e o papel das redes de sociabilidade digitais ao longo dos dias e das semanas que se seguiram aos eventos. Quando considerado o conjunto de registros, nota-se a necessidade de se relativizar a centralidade de instituições como os Arquivos Nacionais e de se repensar a conceptualização de arquivos para entende-los como dados a serem capturados em larguíssima escala.

Para além dessas considerações na Arquivologia, a História tem se dedicado à reflexão sobre como superar o tradicional método de leitura direta e próxima do documento histórico para produzir as narrativas sobre o passado. Ainda que o rito de construção do caminho do historiador para a compreensão do passado (a ida ao Arquivo, o domínio dos catálogos de referência e busca, o entendimento das coleções e a leitura dos textos palavra por palavra) pareça estar comprometida pela explosão no número de documentos, surgem novos recursos e protocolos que, na verdade, preservam a possibilidade de se dominar todo um acervo e de se ler cada documento. A questão passa a ser redefinir o que se entende por "leitura", como com a proposta de Franco Moretti de "leitura distante".

Fazendo menção a como historiadores do porte de Marc Bloch, Fernand Braudel, Pierre Renouvin e Immanuel Wallerstein construíram a acentuada densidade de seus trabalhos, e preocupado com a dificuldade de reprodução de tais métodos nos estudos literários, sobretudo com a relação entre análise e síntese, Moretti desenvolveu uma reflexão para trabalhar com enormes quantidades de textos na Literatura que, agora, pode retribuir a História pelo aprendizado conquistado. Se a escala das fontes mobilizadas por

[27] Considera-se, neste texto, "corpus" (ou "corpora" no plural) como a reunião de textos a serem examinados em conjunto em determinada pesquisa ou para determinado fim.

Bloch no início do século XX impressionaram Moretti (que cita a seguinte frase do historiador francês para ilustrar seu espanto: "anos de análise para um dia de síntese"), a explosão digital de documentos textuais hoje certamente levaria os mesmos historiadores a décadas, não anos, de trabalho manual para produzir o mesmo dia de síntese. A experiência dos estudos literários, então, mostra-se pertinente na reflexão aqui desenvolvida por nos oferecer uma alternativa ao que seria a acima mencionada leitura pormenorizada de um documento, linha por linha, palavra por palavra ou, ainda, a leitura, nesses termos, de apenas um ou de apenas uma minúscula amostra de documentos disponíveis.

Diante desse quadro, apresenta-se logo a questão fundamental: sem poder conhecer, com antecedência, qual texto é certamente relevante para a pesquisa, como decidir, entre milhares ou milhões, quais documentos analisar? Como, então, não ler para, afinal, ler? Se a leitura direta do texto deixar de ser feita, não seria tal escolha, para um pesquisador, um ultraje depois de décadas valorizando a leitura direta de documentos?[28]

Diante da nova escala, na verdade, não se estaria necessariamente eliminando o exame direto de documentos textuais, mas apenas se recorrendo a ferramentas que permitam retomar o controle da escolha sobre quais documentos ler ou ainda de como ler. Enquanto Moretti justifica uma leitura distante pela seleção de unidades de análise menores ou muito maiores do que os limites de um único texto literário (como "recursos literários, temas, figuras de linguagem – ou gêneros"), que podem provocar o abandono do texto original como unidade de análise em nome da compreensão de um "sistema" ou de uma tradição literária[29], o historiador (ou todo aquele interessado em análise de grandes quantidades de texto) pode fazer o mesmo tanto para delimitar seu corpus quanto para analisá-lo. Se, antes, o historiador construía seu corpus a partir da revisão bibliográfica, da consulta a catálogos, da visita a Arquivos, então preparando o conjunto orgânico a ser trabalhado, no futuro, diante de uma escala de informação disponível que impede essas ações, ele adicionará a esse processo recursos que o permitam continuar construindo crítica e conscientemente o conjunto a ser analisado. Trata-se de um exame a partir de uma perspectiva diferente, que tanto permite conservar a observação clássica

[28] MORETTI, Franco. **Distant Reading**. Londres: Verso, 2013, p. 47-48.

[29] MORETTI, Franco. Op. cit., p. 49.

como identificar elementos que antes não eram visíveis. Como notaremos abaixo, novas ferramentas poderão permitir a esse historiador agrupar tematicamente textos de forma automática e estatisticamente relevante, ordená-los de acordo com novos critérios e parâmetros e gerar visualizações completamente originais.

Meios de organizar grandes quantidades de texto não são novidades. A questão passa a ser a formação de parcerias com a Ciência da Computação para desenvolver as ferramentas mais adaptadas à área de aplicação pretendida. Faz-se necessário um diálogo frutífero e compassado para determinar a qualidade do corpus examinado, os parâmetros de processamento, as singularidades do tratamento dos textos e a capacidade de avaliação[30]. Trata-se, entretanto, de uma interação entre disciplinas pouco evidente, demandando uma correta avaliação e reconhecimento do comprometimento na construção e evolução de parcerias, por exemplo, entre o historiador, o cientista da informação, o arquivista, o linguista, o analista de dados e o responsável pelo desenvolvimento de sistemas inteligentes.

Finalmente, pode-se perguntar se não há perdas nesses novos processos de investigação, particularmente na mencionada "leitura distante". Moretti não foge à questão e não pretende dar uma resposta definitiva e totalizante, ressaltando que não há substituição de formas de análise, apenas complementariedade. Interessado em compreender sistemas, assim como propuseram entender o passado de forma sistêmica e total Bloch, Braudel, Renouvin e Wallerstein, Moretti defende que perdas podem acontecer, mas que há enorme potencial para que os ganhos as superem[31].

Algumas ferramentas e seus potenciais

Pelo já apresentado, se a organização de milhares, por vezes milhões, de documentos impõe novos desafios e se sua análise se encontra em risco, quais recursos hoje podem ser considerados para aperfeiçoar ou mesmo inovar metodologicamente e oferecer soluções? Sem que se deixe de considerar grandes quantidades de dados numéricos, nem mesmo o cada vez mais presente registro audiovisual, neste texto, temos

[30] Interessantes exemplos podem ser consultados e explorados no website do projeto *History Lab* (History Lab. Disponível em <www.history-lab.org>. Acessado em: 12 Nov. 2019).

[31] MORETTI, Franco. Op. cit., p. 49.

considerado a análise da linguagem humana como insumo central para pesquisas, sobretudo em sua manifestação textual-discursiva.

Nesse quadro, esforços inovadores e recentes indicam a mineração de textos e o processamento de linguagem natural como caminhos possíveis a partir de exercícios como: Modelagem de Tópicos, Desambiguação Semântica, Contagem de Sequência de Palavras, Análise de Tráfego e de Sentimento, Atribuição de Autoria e Análise de Rede. Com eles, é possível vislumbrar novos níveis de decodificação automática de milhões de sinais gráficos e a manutenção da capacidade humana em continuar a interrogar corpora de textos nas escalas que conhecemos hoje.

Modelagem de Tópicos

De forma muito comum, diante de uma coleção enorme de documentos textuais não estruturados, o principal meio de acesso é começar a digitar palavras-chave em um sistema de buscas simplificado, caso ele exista. Entretanto, reflexões sobre pesquisas dentro do mundo jurídico, por exemplo, já demonstraram que, geralmente, estamos errados ao supor que sabemos quais palavras são realmente "chave", o que leva a perdas médias de cerca de 80% dos documentos relevantes para o que buscamos[32]. Mesmo quando encontramos documentos, pode ser difícil ou impossível reconstruir seu contexto ou significado original. As preocupações, então, são de encontrar uma maneira alternativa de entender o todo e de começar a identificar os grupos de documentos relacionados que são mais representativos dos temas (ou tópicos) nos quais estamos interessados. Existem maneiras potentes e fáceis de conduzir essa exploração e organizar automaticamente milhares ou milhões de documentos em grupos coesos. Para tanto, é possível recorrer a modelos de avaliação de ocorrências de palavras em documentos, como o algoritmo *latent Dirichlet allocation*[33], que pertence a uma categoria geral de modelos de variáveis latentes que inferem tópicos de documentos usando uma abordagem dita de "bag-of-words" (cesto

[32] PECK, Andrew. search, forward Will manual document review and keyword searches be replaced by computer-assisted coding?, **Law Technology News**, Outubro 2011. Disponível em <https://openairblog.files.wordpress.com/2011/11/peck-search-forward.pdf >. Acessado em: 12 Nov. 2019.

[33] BLEI, David M., NG, Andrew Y., JORDAN, Michael I. Latent dirichlet allocation, **Journal of Machine Learning Research**, n. 3, 2003, p. 993–1022

de palavras). Esse método trata cada documento como contendo aleatoriamente tópicos e, cada tópico, como uma distribuição de palavras que pode ser identificada através de uma análise estatística e probabilística. Ao produzir uma análise simultânea de milhares ou milhões de documentos, os resultados serão a criação automática de agrupamentos de documentos com mais afinidades entre si.

A contribuição do pesquisador da área de aplicação (por exemplo, de um historiador) para a preparação de tal processamento é fundamental para que parâmetros como eliminação de palavras que não oferecem singularidade aos documentos, elaboração de listagens de termos técnicos de controle e número de grupos de documentos a serem gerados sejam definidos. Finalmente, diante dos resultados, faz-se novamente vital a presença desse mesmo pesquisador junto à equipe de desenvolvimento para que ele possa determinar quais grupos são relevantes ou não dependendo dos objetivos da pesquisa e para que ele examine amostras de documentos para cada tópico gerado, validando sua coerência temática e os denominando. Considerando que os algoritmos trabalham apenas quantitativamente, a intervenção dos especialistas é essencial para descartar tópicos que não sejam considerados significativos para a área de aplicação. Trata-se, então, de um exercício híbrido, de avaliação quantitativa e interpretativa, com um resultado final de um conjunto de tópicos com curadoria e pronto para ser mais explorado. Os resultados podem ser úteis, por exemplo, para priorizar documentos que mereçam análise ou ainda para exercícios de predição[34].

O potencial para se trabalhar com textos históricos (ou quaisquer outros grandes conjuntos de texto) é imenso e pode, ainda, ajudar os cientistas da computação em suas próprias missões de desenvolvimento de sistemas mais sofisticados, sobretudo quando o objetivo é procurar as estruturas intelectuais dos escritos. Finalmente, a conjugação de modelagem com segmentação do tempo ou do tipo documental pode permitir, por exemplo, a descoberta de como conceitos evoluem ou ainda mostrar que tipo de questões são tratadas em documentos secretos e não secretos.

[34] Risi, J., Sharma, A., Shah, R. et al. Predicting history, **Nature Human Behavior**, n.3, 2019, p. 906–912.

Desambiguação semântica

Em um exercício de exploração de grandes conjuntos textuais, a análise discursiva automática, mais particularmente a Desambiguação de Palavras (ou ainda a Extração de Entidades), apresenta-se como uma ferramenta interessante por poder oferecer, como resultado, identificação automática de tipos semânticos como pessoas, locais e tempo. Trata-se de um exercício complementar ao da modelagem de tópicos, sobretudo por permitir expor, automaticamente, os objetivos e a organização do texto, valorizando sua análise linguística, e não simplesmente estatística e probabilística.

Ferramentas como o *parser* Palavras[35] identificam vocábulos e conduzem uma análise morfossintática dos mesmos, produzindo uma segmentação de textos em larguíssima escala, resultando em unidades que contenham uma ideia ou um conceito básico. Cada uma dessas unidades, então, recebe marcações sobre sua classe morfológica e sobre seu papel sintático. Em um corpus onde se esperam, por exemplo, identificar nomes de pessoas e países, faz-se necessário, também, o uso de bibliotecas (listas de referência e controle) que contenham cada um desses termos para que a desambiguação possa ser completada (para países, uma possibilidade seria utilizar a listagem dos atuais membros das Nações Unidas ou, para nomes de pessoas, uma listagem baseada em verbetes de dicionários biográficos). Com tal exercício seria possível, por exemplo, identificar no corpus quando a expressão "Getúlio Vargas" se refere a pessoa, ou quando se refere a logradouro público ou ainda a alguma instituição (como o nome de uma escola).

Ainda que o exercício possa parecer rudimentar à primeira vista, ou que se apresente apenas como uma etapa de uma análise retórica e discursiva do texto, nota-se um enorme potencial para sofisticar buscas em conjuntos gigantescos de textos. A trajetória da Hemeroteca Digital da Biblioteca Nacional (HDBN), por exemplo, uma iniciativa de digitalização de jornais e revistas antigos e de seu tratamento por reconhecimento óptico de caracteres, poderia ganhar muito caso caminhe nessa direção. Desde 2012, a HDBN passou a permitir a interessados buscar e recuperar informações no conteúdo dessas publicações de uma forma sem precedentes. No âmbito da pesquisa histórica, tratou-se de um novo e preciosíssimo recurso, trazendo enorme impacto para a

[35] BICK, Eckhard. **The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. Aarhus University Press, 2000.

condução de pesquisas documentais mais complexas. O mecanismo de busca permite acionar três parâmetros: local de publicação, período de publicação e título do periódico, oferecendo ainda a possibilidade de combinar uma dessas informações com palavras-chave. Entretanto, um maior refinamento, como através de resultados de desambiguação, ainda não é possível. Um usuário interessado em explorar as 15 milhões de páginas já digitalizadas na HDBN, ao fazer uma busca com a palavra "Vargas" em periódicos publicados apenas no Rio de Janeiro, encontrará 1.136.794 ocorrências dentre as quais, para além da referência à pessoa do antigo Presidente, estarão também menções a homônimos, instituições, avenidas e ruas[36].

Contagem de sequência de palavras

Ainda que não haja conversão fácil de palavras em números ou dados para, por exemplo, evocarmos todo o conhecimento da já veterana História Serial a fim de produzir um estudo quantitativo de grandes conjuntos textuais, há meios de realizar diversas modalidades de contagem sofisticada de palavras explorando simultaneamente recursos desenvolvidos pela Cartografia Digital ou pela Linguística Computacional.

Combinações de ferramentas que permitam uma contagem automática de unidades linguísticas (palavras) levando em conta possíveis irregularidades dos elementos que constituem o corpus (como variação do volume do conteúdo de jornais ou de e-mails), mas também suas localizações (como lugar da impressão ou da emissão) e estimativas quanto à ocorrência de palavras com relação a outras palavras do texto podem permitir entender permanências e mudanças. Em um arquivo que inclua milhões de telegramas diplomáticos, por exemplo, esses exercícios podem revelar intensidades e prioridades de uma atividade diplomática ao longo do tempo e em diferentes espaços conjugados ou comparados. Essa combinação pode se transformar, finalmente, em uma maneira visualmente intuitiva de rastrear ideias e conceitos à medida que evoluem.[37]

[36] Hemeroteca Digital da Biblioteca Nacional. Disponível em <http://bndigital.bn.gov.br/hemeroteca-digital/>. Acessado em 12 Nov. 2019.

[37] Para uma apresentação pormenorizada das ferramentas trabalhadas pela Linguística Computacional, ver: FERREIRA, Marcelo e LOPES, Marcos. Linguística Computacional, FIORIN, J.L. (org.). **Novos Caminhos da Linguística**. São Paulo: Contexto, 2017, p. 195-214. Para um exemplo da articulação entre contagem e

O processo de codificação pode ser entrelaçado com bases de dados ou corpus conforme o desejo do pesquisador de modo a realizar funções específicas. Assim, é possível detectar padrões e desenvolvimentos nos textos, comparar a frequência de palavras e relacioná-las com aquelas que a acompanham de maneira objetiva e automática. A possibilidade de formar sequências de conjuntos de palavras permite aumentar o potencial de análise, propiciando exames de expressões, nomes compostos, nomes de instituições etc.[38]

As possibilidades de se criar visualizações quanto à distribuição temporal ou espacial dos termos permitem uma análise para além do nível de atividade, por exemplo, de uma rede de interesse. Cria-se, assim, uma possibilidade de se rastrear e mapear interesses em jogo, de refinar buscas isolando lugares, instituições ou pessoas quando tais dados se encontram inicialmente mergulhados em um universo de documentos não estruturados. Em determinado fluxo de comunicações, como telegramas diplomáticos ou e-mails, faz-se então possível medir a frequência com que se usa, por exemplo, o termo "terrorismo" em comunicações secretas ou não, avaliando quando a questão foi prioritária. As ferramentas podem também ser aplicadas em áreas de estudos como probabilidade, teoria da comunicação, tradução, verificação e correção ortográfica, entre outras, além de possibilitarem a recuperação de informação (como para encontrar documentos e bancos de dados com base em palavras-chave e metadados).

Análise de Tráfego e de Sentimento

Para alguém interessado em analisar relações entre agrupamentos humanos e seus meios de comunicação e relacionamento, mas que se depara com milhões de registros textuais como cartas, e-mails, telegramas diplomáticos ou ainda mensagens transmitidas via mídias sociais, a análise dos fluxos dessas comunicações pode permitir a manutenção do controle sobre o corpus. A escolha das unidades de tempo a serem consideradas, como horas, dias, meses ou anos, dependerá dos tempos históricos em consideração e do

georreferenciamento, ver: BLEVINS, Cameron. Space, Nation, and the Triumph of Region: A View of the World from Houston, **Journal of American History**, vol. 101, n. 1, 2014, p. 122–147.

[38] Um exemplo de uso de combinações para análise textual é MORETTI, Franco e PESTRE, Dominique. Bankspeak: the language of World Bank reports, **New Left Review**, vol. 92, n.2, 2015, p. 75-99.

propósito da pesquisa, com a ressalva de que se faz interessante ir além da simples detecção de alterações no ritmo das comunicações.

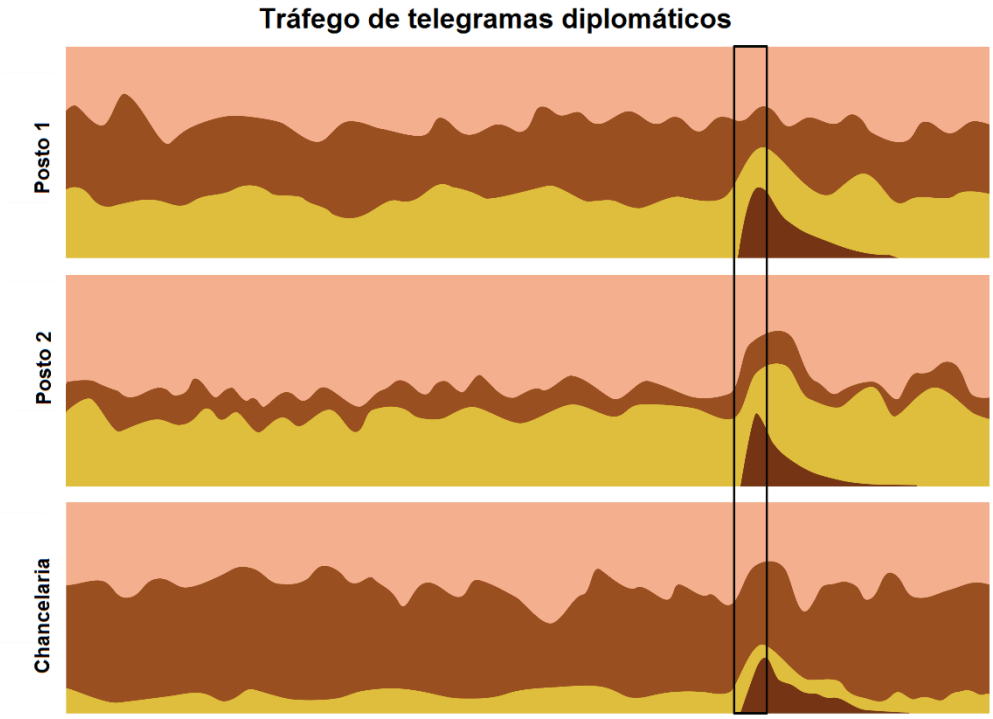

Imagem 2: Nesta simulação, o eixo y representa as proporções acumuladas de telegramas sobre vários tópicos, enquanto o eixo x representa o tempo. O posto diplomático 1, 2 e a Chancelaria apresentam assuntos diplomáticos sobre os quais costumam normalmente enviar a mensagem. Quando um evento excepcional ocorre (destacado na imagem), os telegramas são alterados para cobrir o evento antes de retornar à normalidade.

Trata-se de um exercício com potencial para revelar crises e suas diferentes percepções ou, ainda, como processos de tomada de decisão são afetados. Um chefe de missão diplomática, por exemplo, pode intensificar ou alterar o conteúdo das comunicações telegráficas com a sede ministerial por entender se aproximar um golpe de

Estado na localidade onde se encontra. O fluxo das respostas da respectiva Chancelaria, entretanto, pode não se alterar, levantando questões sobre diferenças de percepção (ou sobre a importância excepcional e, talvez, exagerada que o embaixador parece dar ao que ele próprio entende descrever como relevante). Em uma reflexão pioneira sobre a personalidade de atores históricos, Jean-Baptiste Duroselle já dizia que os tomadores de decisão "afoitos por glória (...) são numerosos na história e perigosos para a paz"[39]. Um dos desafios de tal exercício, então, encontra-se na determinação das estruturas regulares de comunicação e, por consequência, na determinação de irrupções nos textos examinados, como indica Jon Kleinberg[40]. Adicionalmente, pode-se produzir um estudo também automático, mas mais apurado dos conteúdos dos fluxos de comunicação e da linguagem utilizada, como uma análise de sentimentos, opiniões e avaliações. Ferramentas como o *Quanteda*[41] permitem a conversão de palavras em sinais positivos e negativos, dentre outras possibilidades, aperfeiçoando a análise de fluxos de comunicação, podendo permitir até mesmo exercícios de detecção e caracterização de eventos históricos, que podem interessar tanto historiadores como cientistas políticos e jornalistas.

Atribuição de Autoria

Em 2016, Felipe Botelho Coelho lançou *Sátiras e outras subversões*, um livro com 164 textos de Lima Barreto que, até então, não tinham a autoria identificada. Assim como no tempo do escritor carioca, em diversos momentos foi comum a prática de ocultar a identificação a fim de evitar retaliações por atritos com pares ou com poderosos. Para revelar a autoria nesse caso particular, Coelho não recorreu, porém, a técnicas computacionais e enfrentou um árduo trabalho de cinco anos para dominar em profundidade tanto a obra de Lima Barreto como o tempo em que ele viveu e que tentou

[39] RENOUVIN, Pierre e DUROSELLE, Jean-Baptiste. **Introdução à História das Relações Internacionais**. São Pauo: Difel, 1967 (1964), p. 326.

[40] KLEINBERG, Jon. Bursty and Hierarchical Structure in Streams, **Data Mining and Knowledge Discovery**, n. 7, 2003, p. 373–97.

[41] BENOIT, Kenneth et al.., (2018). quanteda: An R package for the quantitative analysis of textual data. **Journal of Open Source Software**, vol. 3, n. 30, 2018.

marginaliza-lo, os meios de comunicação e os pseudônimos que utilizou e os recursos estéticos e o tipo de intervenção literária que o singularizaram. Tudo isso, a fim de que pudesse traçar paralelos e atribuir quase que artesanalmente a autoria de textos publicados há cem anos. Tratou-se de um intrincado trabalho que, segundo o próprio Coelho, ainda pode ter deixado para trás inúmeros outros textos não identificados[42].

Longe de ser uma técnica de análise de texto original, a Atribuição de Autoria aparece hoje como inovadora, na verdade, quanto à adoção de técnicas computacionais como o aprendizado de máquina para alcançar mais rapidez e precisão nos resultados, sobretudo quanto aos seguintes desafios: (i) quando há ausência de prováveis nomes de autor para os textos, o que demanda a criação de um perfil o mais preciso possível, (ii) quando existem vários nomes de autor em potencial ou, ainda, (iii) quando existe apenas um possível nome, demandando a chamada "verificação de autoria".

Para o primeiro desafio, experiências recentes têm utilizado um corpus extenso e variado para indicar, através dos textos examinados, particularidades como gênero, idade, língua materna, personalidade, entre outros, buscando enquadrar o autor anônimo segundo suas características pessoais. Já para o segundo desafio, deve haver recurso a corpora significativos de trabalhos de todos os prováveis autores, a fim de que se possa definir a autoria por meio da comparação do texto alvo com os padrões e as peculiaridades da escrita de cada um deles. Já o terceiro, demanda que se encontrem previamente, nos textos de autoria já determinada, padrões e características da escrita do autor (como frequência de palavras, marcas de pontuação, tamanho médio de sentenças, riqueza de vocabulário, escolha de sinônimos e construção sintática, entre outros), de modo a verificar suas existências no texto em exame. Como indicam Daniela Witten e Matthew Jockers, trabalhando com a identificação de textos dos chamados "pais fundadores" dos Estados Unidos (os *Federalist Papers*), para além das ferramentas computacionais empregadas mais comumente em atribuições de autoria (como o *support vector machine*), é possível recorrer a diversos métodos de aprendizado de máquina (como *nearest shrunken centroids*) para também vencer os desafios mencionados[43].

[42] COELHO, Felipe Botelho (org.). **Sátiras e outras subversões: textos inéditos**. São Paulo: Penguim Classics Companhia das Letras, 2016, p. 11-75.

[43] JOCKERS, Matthew L. e WITTEN, Daniela M. A Comparative Study of Machine Learning Methods for Authorship Attribution, **Literary and Linguistic Computing**, n. 25, 2010, p. 215–23. Ver, também: KOPPEL, Moshe, SCHLER, Jonathan e ARGAMON, Shlomo. Computational methods in authorship attribution.

Análise de Rede

Mais do que meio de visualização de relações entre atores sob exame, a análise de rede pode ser utilizada como ferramenta na pesquisa com textos. Partindo de largas coleções de documentos, faz-se possível reconstituir, por exemplo, redes de relacionamento e sociabilidade e levantar questionamentos sobre a centralidade e importância de um ator, ou ainda sobre a circulação de informações através de uma estrutura burocrática. O acesso a dados de trocas de mensagens de mídias sociais pode prover uma análise ainda mais sofisticada de redes de sociabilidade, demonstrando conexões que, por exemplo, não são oficiais (redes informais) diante de um processo de tomada de decisão. Uma outra possibilidade é entender o impacto de mudanças fundamentais ou crises (como golpes de Estado ou revoluções) para alterações nas redes de sociabilidade.

A análise de rede, ainda mais do que outros meios de visualização de dados (como gráficos, nuvens de palavras etc), possibilita uma leitura diferente de um corpus textual. Através de ferramentas como o *Gephi*[44], a teia pode ser formada com base em temas, palavras-chave, pessoas e lugares e pode ter sua estrutura definida por meio de critérios de centralidade pré-estabelecidos, esclarecendo, assim, a posição na rede dos objetos (chamados de "nós") estudados[45]. Esses podem ser analisados, então, por sua quantidade de conexões com outros nós ("centralidade de grau"), por sua proximidade com a totalidade da rede ("centralidade de proximidade"), por seu papel de nó "ponte", conectando outros nós ("centralidade de intermediação") e por sua quantidade de conexões com outros nós, ponderando, dessa vez, a qualidade de cada conexão – com a

Journal of the American Society for information Science and Technology, v. 60, n. 1, 2009, p. 14-21 e GRIEVE, Jack. Quantitative authorship attribution: An evaluation of techniques, **Literary and linguistic computing**, v. 22, n. 3, 2007, p. 251-270.

[44] Gephi. Disponível em <https://gephi.org/>. Acessado em 12 Nov. 2019.

[45] ATTRIDE-STIRLING, Jennifer. Thematic networks: an analytic tool for qualitative research, **Qualitative research**, v. 1, n. 3, 2001, p. 385-405.

alta qualidade significando que o nó é conectado com outros nós que também possuem alta qualidade ("centralidade de prestígio")[46].

O pesquisador é capaz, assim, de montar visualizações sobre as conexões entre as personagens que analisa, sobre a intensidade das relações diplomáticas entre Estados, por exemplo, ou ainda sobre a utilização de linhas de metrô e trem e mesmo sobre as relações temáticas entre textos de um corpus.

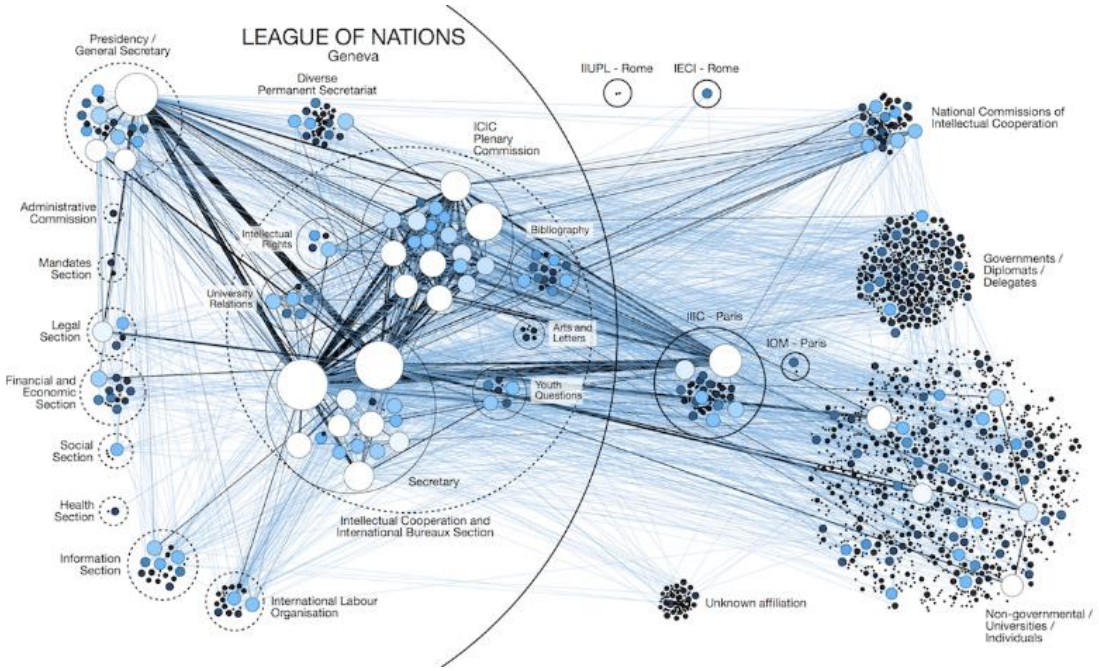

Imagem 3: Visualização criada pelo historiador Martin Grandjean da rede de cerca de 1700 pessoas (com dezenas de milhares de relacionamentos) ligadas ao Comitê Internacional de Cooperação Intelectual (1919-1927)[47]

[46] GRANDJEAN, Martin. Introduction à la visualisation de données: l'analyse de réseau en histoire, **Geschichte und Informatik, Chronos**, 2015, p. 109-128. Para mais informações sobre o processo de modelagem, ver GRANDJEAN, Martin. GEPHI - Introduction to Network Analysis and Visualization, **Martin Grandjean: Digital Humanities, Data Visualisation, Network Analysis**, 2015. Disponível em: <http://www.martingrandjean.ch/gephi-introduction/>. Acesso em: 24 Out. 2019.

[47] GRANDJEAN, Martin. Intellectual Cooperation: multi-level network analysis of an international organization, 15/12/2014. Disponível em <http://www.martingrandjean.ch/intellectual-cooperation-multi-level-network-analysis/>. Acesso em : 29 Out. 2019.

Conclusão

A digitalização dos registros textuais da experiência humana pode se apresentar inicialmente como redenção para o problema da preservação. Como visto, entretanto, as novas escalas de documentos preservados criam novos desafios. Na verdade, antes mesmo dessas questões surgirem, seria preciso discutir a paradoxal eliminação de vestígios que a própria digitalização pode produzir.

William Turkel, por exemplo, nos lembra desse problema ao mencionar como historiadores da saúde podem procurar, em cartas, tanto o conteúdo textual como o aromático. Ele cita um caso em particular de cartas do século XVII que, na época, para conter uma epidemia de cólera, foram banhadas em vinagre. Um historiador que constate o odor de vinagre e faça a correlação entre a data e o lugar em que as cartas foram escritas pode retraçar o caminho e a intensidade da epidemia [48]. O exemplo apenas nos indica que o passado sempre será tirânico para com aquele que o procure enfrentar. O recurso a sistemas inteligentes, porém, pode ser de imensa valia, sobretudo quando feito em consonância com as singularidades das áreas de aplicação como a História, a Sociologia, o Direito, a Comunicação e diversas outras.

Alguns leitores ainda podem argumentar, de forma oportuna, que diversas das ferramentas e recursos aqui brevemente apresentados encontram-se distantes da realidade de muitas instituições arquivísticas nacionais, mais comumente às voltas com problemas de infraestrutura e falta de recursos. Não obstante, entendemos que eles podem servir ao menos de inspiração para mudanças de perspectiva nas formas como se pensa o trabalho cotidiano com documentos textuais e para planejamentos futuros. Talvez possam ainda contribuir para transformar algumas práticas, permitindo um mais fecundo aproveitamento da implementação de sistemas inteligentes quando ela for materialmente possível. Nesse sentido, por exemplo, em instituições conduzindo a digitalização de seus acervos, ainda que o ritmo possa se reduzir, ao dirigir o máximo investimento possível para a qualidade do processo e do resultado (utilizando peritos e equipamentos apropriados) garante-se uma exploração ainda mais elevada do material no futuro, além de se evitar uma eventual repetição do processo, certamente dispendiosa. Em um outro plano, a promoção de diálogos multidisciplinares, ao alcance hoje da simples escolha de

[48] TURKEL, Dans William J. Intervention: Hacking history, from analogue to digital and back again, **Rethinking History**, vol. 15, n. 2, 2011, p. 287-296.

profissionais envolvidos com todo e qualquer aspecto da digitalização de documentos textuais, pode gerar uma cultura institucional comum que fundamente, no futuro, a busca de soluções para um fenômeno que tem abolido as fronteiras entre disciplinas tão diversas como a Ciência da Informação, a Arquivologia e a Ciência da Computação. Mesmo para instituições arquivísticas que ainda não preservam acervos digitais e que enfrentam graves dificuldades materiais, muitas vezes tendo como maior luta a mera manutenção de suas salas de consulta abertas ao público, algumas simples iniciativas podem render grandes frutos. Em uma realidade na qual a grande maioria dos pesquisadores visita acervos, faz buscas e captura documentos utilizando câmeras fotográficas digitais[49], pensar em oferecer rápidas oficinais de treinamento de manipulação de tais equipamentos (e dos próprios textos a serem fotografados) pode permitir uma prestação de serviço de muita qualidade para o interessado, além de contribuir para uma melhor preservação dos documentos.

Finalmente, quanto ao trabalho com texto em meio à disponibilidade de ferramentas e capacidades computacionais inéditas, é preciso alertar para o fato de que estamos em um momento diferente daquele da História Serial de meados do século XX. Ao mesmo tempo em que a preocupação não são os dados, mas a linguagem humana em registros textuais-discursivos, ainda necessitamos de mais recursos que trabalhem com a Língua Portuguesa e que possam fazer análises não somente morfológicas ou sintáticas para a exploração de padrões léxico-gramaticais, como as aqui expostas demonstraram, mas que sigam adiante e produzam análises semânticas e interpretativas para, finalmente, tomar decisões e sanar, por exemplo, a dificuldade em administrar (e liberar) os documentos secretos mantidos pelo Estado.

Mais do que alcançar predição ou previsão, estaríamos rumando em direção a uma dinâmica de criação de exercícios que reproduzam a capacidade humana de interpretação. Trata-se de percorrer um caminho que nos permita especular sobre a invenção de soluções automatizadas para também produzir tramas e narrativas a partir da organização de

[49] Desenvolvendo um singular levantamento no Canadá, o professor Ian Milligan apurou que entre 90% e 95% dos historiadores usa câmeras digitais para fotografar documentos quando realizando pesquisa em Arquivos. Milligan descobriu, nessa mesma pesquisa, que 96% dos historiadores não possuem nenhuma formação para trabalhar com fotos digitais, mas que, ao menos metade, estaria disposta a recebê-la (MILLIGAN, Ian. Becoming a Desk(top) Profession: Digital Photography and the Changing Landscape of Archival Research. **American Historical Association Annual Meeting**. Nova York, 2020).

eventos aparentemente caóticos do passado, que se imaginava apenas poder emergir do gênio humano. Como lembra o filósofo Daniel Dennett, porém, faz-se vital um alerta: não devemos percorrer leviana ou inconscientemente essas vias, ainda menos celebrar a perda de controle de diversos aspectos de nosso destino que levamos séculos para conquistar[50].

[50] DENNETT, Daniel. The Singularity – an Urban Legend?, BROCKMAN, John (org.). **What to Think About Machines That Think?** Nova York: Harper Perennial, 2015, p. 85-88